# Dr. SAJAN KUMAR

(778)-558-7592 skumarudel@gmail.com skumarudel Sajan-Kumar Personal website

## PROFESSIONAL SUMMARY

Highly analytical Data Engineer / Senior Data Scientist with over 10 years of experience architecting scalable data pipelines, processing large-scale unstructured datasets (2+ TB), and building production-grade AI infrastructure. Expertise in pythonic ETL/ELT pipeline development, multi-source data integration, advanced statistical modeling, machine learning, vector database architecture, and AWS cloud environments. Proven track record of translating complex, chaotic streaming and sensor data into optimized, enterprise-ready data environments.

## Technical Skills

- **Data Engineering:** ETL/ELT Pipelines, Data Ingestion, Data Normalization, Schema Design, Data Validation, High-Throughput Processing
- **Languages & Core Frameworks:** SQL, Python (Pandas, NumPy), Linux Shell Scripting
- **Cloud & Dev Tools:** AWS (S3, Lambda, SageMaker, Bedrock), Docker, Git, HPC (Slurm), Airflow, Snowflake
- **Machine Learning & AI:** Scikit-learn, PyTorch, Statistical modeling, Dimensionality Reduction (PCA, UMAP), Clustering (HDBSCAN), Regression and Classification (Decision trees), time series forcasting (ARIMA, LSTM)
- **Vector DBs & LLM Systems:** ChromaDB, Semantic Search Architecture, RAG Pipelines and LLM integration

## Experience

### Mercor | Remote, Canada

**AI Data & Model Evaluator (Physics & Analytics Consultant)** 2025 – Present

- Evaluate and optimize LLM training data and code outputs for technical accuracy, execution logic, and rigorous data validation.
- Implement evaluation frameworks to ensure production-grade data reasoning meets strict enterprise standards for regulatory compliance and AI safety.

### University of Maryland | College Park, USA

**Senior Data Engineer & Analytics Specialist** 2022 – Present

- **Data Infrastructure Integration:** Engineered a custom data-unification plugin to merge, transform, and normalize massive, heterogeneous streaming and sensor datasets from multiple independent observatories into a unified analytics platform. ***Impact:*** Improving data accessibility and precise modeling.
- **Production RAG Architecture:** Designed and developed a scalable semantic retrieval layer using SPECTER2 dense embeddings and ChromaDB to efficiently index and retrieve relevant documents from 50,000+ unstructured technical documents. ***Impact:*** Significantly improving search relevance over keyword-based approaches.
- **Context-Aware LLM Pipeline:** Implemented a reranking and Retrieval-Augmented Generation (RAG) workflow that combined retrieved context with LLM inference to generate accurate, context-aware responses, and deployed an interactive Streamlit application. ***Impact:*** Efficient retrieval and synthesize of relevant scientific knowledge from large-scale document collections.
- **Production Engineering:** Engineered modular, Dockerized, version-controlled data and AI pipelines with automated validation and reproducible workflows. ***Impact:*** Improved reliability, and simplified maintenance and collaboration across the research team.

**McGill University & University of Delaware | Montreal, Canada / Delaware, USA**

**Data Pipeline & Analytics Engineer** 2018 – 2022

- **Big Data Processing (2 TB+):** Managed and optimized a big data infrastructure pipeline processing 2+ Terabytes of spatiotemporal observational data, cutting through data noise to surface rare-event detections. ***Impact:*** Improved detection sensitivity by a factor of 2 over prior methods (Paper link).
- **Synthetic Data Generation:** Programmed high-throughput Monte Carlo data generation pipelines to generate 100+ million synthetic events that accurately simulate complex environmental and sensor inputs. ***Impact:*** Providing highly accurate synthetic baseline datasets for model training.
- **Performance Optimization:** Engineered scalable feature extraction and preprocessing pipelines for large image datasets, and trained optimized Random Forest and Boosted Decision Tree models. ***Impact:*** Achieved 90% background rejection while maintaining over 95% signal efficiency, enabling more accurate identification of rare scientific events and improving downstream analysis.

## Project Highlights

**ArXiv-RAG: Intelligent Literature Retrieval System [Manuscript in preparation]**

- Architected a complete data pipeline that ingests, cleans, and indexes thousands of scientific documents, enabling scalable retrieval over large unstructured corpora.
- Implemented a RAG pipeline that integrates reranking and LLM-based synthesis (Gemma4 model) to generate context-aware responses from retrieved scientific literature.

## Leadership

**Technical Committee Lead & Program Chair | VERITAS Collaboration** 2020-2024

- **Project Management:** Directed a cross-functional international team of 100+ scientists and engineers, developing scientific proposals, and defining project future directions.
- **Resource Allocation:** Governed data collection priority roadmaps for approximately 40% of the entire organization's telescope array operations.
- **Pipeline Delivery:** Streamlined the end-to-end internal lifecycle from raw data ingestion to final technical validation and publication.
- **Mentoring:** Provided technical and career mentorship to 5+ graduate students, guiding validation testing, pipeline development, and codebase integration, culminating in successful PhD defenses and professional growth.
- **Communication:** Delivered 20+ invited and contributed presentations at international conferences and co-authored 50+ peer-reviewed research publications, effectively communicating complex technical and scientific concepts to diverse academic and interdisciplinary audiences.

## Education

**University of Delaware | Newark, USA** 2010-2018

**Ph.D. in Physics & Astronomy (Focus on High-Performance Computational & Statistical Analysis)**

- **Key focus:** Statistical inference, large-scale computational modeling, Linux cluster (HPC) management, big data analysis, scientific writing.

## Awards & Fellowships

- **NASA Guest Investigator Grant ($70,000):** Secured as Principal Investigator to fund advanced computational data analysis.
- **University Dissertation Fellowship ($30,000):** Awarded for high-impact analytical research.
- **iProgress Scholarship:** Awarded by the Helmholtz Alliance for Astroparticle Physics, Germany.